Efficient volume sampling for row/column subset selection

Amit Deshpande
Microsoft Research India
amitdesh@microsoft.com

Luis Rademacher
Computer Science and Engineering
Ohio State University
lrademac@cse.ohio-state.edu

Abstract

We give efficient algorithms for volume sampling, i.e., for picking k-subsets of the rows of any given matrix with probabilities proportional to the squared volumes of the simplices defined by them and the origin (or the squared volumes of the parallelepipeds defined by these subsets of rows). This solves an open problem from the monograph on spectral algorithms by Kannan and Vempala (see Section 7.4 of [15], also implicit in [1, 5]).

Our first algorithm for volume sampling k-subsets of rows from an m-by-n matrix runs in $O(kmn^\omega \log n)$ arithmetic operations and a second variant of it for $(1+\epsilon)$-approximate volume sampling runs in $O(mn \log m \cdot k^2/\epsilon^2 + m \log^\omega m \cdot k^{2\omega+1}/\epsilon^{2\omega} \cdot \log(k\epsilon^{-1} \log m))$ arithmetic operations, which is almost linear in the size of the input (i.e., the number of entries) for small k.

Our efficient volume sampling algorithms imply the following results for low-rank matrix approximation:

1. Given $A \in \mathbb{R}^{m\times n}$, in $O(kmn^\omega \log n)$ arithmetic operations we can find k of its rows such that projecting onto their span gives a $\sqrt{k+1}$-approximation to the matrix of rank k closest to A under the Frobenius norm. This improves the $O(k\sqrt{\log k})$-approximation of Boutsidis, Drineas and Mahoney [1] and matches the lower bound shown in [5]. The method of conditional expectations gives a *deterministic* algorithm with the same complexity. The running time can be improved to $O(mn \log m \cdot k^2/\epsilon^2 + m \log^\omega m \cdot k^{2\omega+1}/\epsilon^{2\omega} \cdot \log(k\epsilon^{-1} \log m))$ at the cost of losing an extra $(1+\epsilon)$ in the approximation factor.
2. The same rows and projection as in the previous point give a $\sqrt{(k+1)(n-k)}$-approximation to the matrix of rank k closest to A under the spectral norm. In this paper, we show an almost matching lower bound of $\sqrt{n}$, even for $k = 1$.

Keywords: volume sampling, low-rank matrix approximation, row/column subset selection

1 Introduction

Volume sampling, i.e., picking k-subsets of the rows of any given matrix with probabilities proportional to the squared volumes of the simplicies defined by them, was introduced in [5] in the context of low-rank approximation of matrices. It is equivalent to sampling k-subsets of $\{1, \dots, m\}$ with probabilities proportional to the corresponding k by k principal minors of any given m by m positive semidefinite matrix.

In the context of low-rank approximation, volume sampling is related to a problem called *row/column-subset selection* [1]. Most large data sets that arise in search, microarray experiments, computer vision, data mining etc. can be thought of as matrices where rows and columns are indexed by objects and features, respectively (or vice versa), and we need to pick a small subset of features that are dominant. For example, while studying gene expression data biologists want a small subset of genes that are responsible for a particular disease. Usual dimension reduction techniques such as principal component analysis (PCA) or random projection fail to do this as they typically output singular vectors or random vectors which are linear combinations of a large number of feature vectors. A recent article by Mahoney and Drineas [18] highlights the limitations of PCA and gives experimental data on practical applications of low-rank approximation based on row/column-subset selection.

2 Row/column-subset selection and volume sampling

While dealing with large matrices in practice, we seek smaller or low-dimensional representations of them which are close to them but can be computed and stored efficiently. A popular notion for low-dimensional representation of matrices is low-rank matrices, and the most popular metrics used to measure the closeness of two matrices are the Frobenius or Hilbert-Schmidt norm (i.e., the square root of the sum of squares of entries of their difference) and the spectral norm (i.e., the largest singular value of their difference). The singular value decomposition (SVD) tells us that any matrix $A \in \mathbb{R}^{m \times n}$ can be written as

$$A = \sum_{i=1}^{m} \sigma_i u_i v_i^T,$$

where $\sigma_1 \geq \ldots \geq \sigma_m \geq 0$, $u_i \in \mathbb{R}^m$ are orthonormal and $v_i \in \mathbb{R}^n$ are orthonormal. Moreover, the nearest rank-k matrix to A, let us call it A_k, under both the Frobenius and the spectral norm, is given by

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T.$$

In other words, the rows of A_k are projections of the rows of A onto $\operatorname{span}(v_i \ : \ 1 \leq i \leq k)$. Because of this, most dimension reduction techniques based on the singular value decomposition, e.g., principal component analysis (PCA), are interpreted as giving v_i's as the *dominant* vectors, which happen to be linear combinations of a large number of the rows or feature vectors of A.

The *row-subset selection* problem we consider in this paper is: Can we pick a k-subset of the rows (or feature vectors) of $A \in \mathbb{R}^{m \times n}$ so that projecting onto their span is almost as good as projecting onto $\operatorname{span}(v_i \ : \ 1 \leq i \leq k)$?

Several row-sampling techniques have been considered in the past as an approximate but faster alternative to the singular value decomposition, in the context of streaming algorithms and large data sets that cannot be stored in random access memory [8, 7, 5]. The first among these is the squared-length sampling of rows introduced by Frieze, Kannan and Vempala [8]. Another sampling scheme due to Drineas, Mahoney and Muthukrishnan [7] uses the singular values and singular vectors to decide the sampling probabilities. Later Deshpande, Rademacher, Vempala and Wang [5] introduced volume sampling as a generalization of squared-length sampling.

Definition 1. *Given $A \in \mathbb{R}^{m \times n}$, volume sampling is defined as picking a k-subset S of $[m]$ with probability proportional to*

$$\det\left(A_S A_S^T\right) = \left(k! \cdot \operatorname{vol}\left(\operatorname{conv}\left(\{\bar{0}\} \cup \{a_i \ : \ i \in S\}\right)\right)\right)^2,$$

where a_i denotes the i-th row of A, $A_S \in \mathbb{R}^{k \times n}$ denotes the row-submatrix of A given by rows with indices $i \in S$, and $\mathrm{conv}\,(\cdot)$ *denotes the convex hull.*

The application of volume sampling to low-rank approximation and, more importantly, to the *row-subset selection* problem, is given by the following theorem shown in [5]. It says that picking a subset of k rows according to volume sampling and projecting all the rows of A onto their span gives a $(k+1)$-approximation to the nearest rank-k matrix to A.

Theorem 2. *[5] Given any $A \in \mathbb{R}^{m \times n}$,*

$$\mathsf{E}\left[\|A - \pi_S(A)\|_F^2\right] \leq (k+1)\,\|A - A_k\|_F^2\,,$$

when S is picked according to volume sampling, $\pi_S(A) \in \mathbb{R}^{m \times n}$ denotes the matrix obtained by projecting all the rows of A onto $\mathrm{span}\,(a_i\,:\, i \in S)$*, and A_k is the matrix of rank k closest to A under the Frobenius norm.*

As we will see later, this easily implies

$$\mathsf{E}\left[\|A - \pi_S(A)\|_2\right] \leq \sqrt{(k+1)(n-k)}\,\|A - A_k\|_2\,.$$

Theorem 2 gives only an existence result for row-subset selection and we also know a matching lower bound that says this is the best we can possibly do.

Theorem 3. *[5] For any $\epsilon > 0$, there exists a matrix $A \in \mathbb{R}^{(k+1) \times k}$ such that picking any k-subset S of its rows gives*

$$\|A - \pi_S(A)\|_F \geq (1-\epsilon)\sqrt{k+1}\,\|A - A_k\|_F\,.$$

However, no efficient algorithm was known for volume sampling prior to this work. An algorithm mentioned in Deshpande and Vempala [6] does $k!$-approximate volume sampling in time $O(kmn)$, which means that plugging it in Theorem 2 can only guarantee $(k+1)!$-approximation instead of $(k+1)$. Finding an efficient algorithm for volume sampling is mentioned as an open problem in the recent monograph on spectral algorithms by Kannan and Vempala (see Section 7.4 of [15]).

Boutsidis, Drineas and Mahoney [1] gave an alternative approach to row-subset selection (without going through volume sampling) and here is a re-statement of the main theorem from their paper which uses columns instead of rows.

Theorem 4. *[1] For any $A \in \mathbb{R}^{m \times n}$, a k-subset S of its rows can be found in time $O\left(\min\{mn^2, m^2n\}\right)$ such that*

$$\begin{aligned}
\|A - \pi_S(A)\|_F &= O(k\sqrt{\log k})\,\|A - A_k\|_F \\
\|A - \pi_S(A)\|_2 &= O\left(k^{3/4}(n-k)^{1/4}\sqrt{\log k}\right)\|A - A_k\|_2\,.
\end{aligned}$$

Row/column-subset selection problem is related to rank-revealing decompositions considered in linear algebra [12, 19], and the previous best algorithmic result for row-subset selection in the spectral norm case was given by a result of Gu and Eisenstat [12] on strong rank-revealing QR decompositions. The following theorem is a direct consequence of [12] as pointed out in [1].

Theorem 5. *Given $A \in \mathbb{R}^{m \times n}$, an integer $k \leq n$ and $f \geq 1$, there exists a k-subset S of the columns of A such that*

$$\left\|A^T - \pi_S(A^T)\right\|_2 \leq \sqrt{1 + f^2 k(n-k)}\, \|A - A_k\|_2 .$$

Moreover, this subset S can be found in time $O\left((m + n \log_f n)n^2\right)$.

In the context of volume sampling, it is interesting to note that Pan [19] has used an idea of picking submatrices of *locally maximum volume* (or determinants) for rank-revealing matrix decompositions. We refer the reader to [19] for details.

The results of Goreinov, Tyrtyshnikov and Zamarashkin [10, 11] on pseudo-skeleton approximations of matrices look at submatrices of maximum determinants as good candidates for row/column-subset selection.

Theorem 6. *[10] If $A \in \mathbb{R}^{m \times n}$ can be written as*

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

where $A_{11} \in \mathbb{R}^{k \times k}$ is the k by k submatrix of A of maximum determinant. Then,

$$\max_{i,j} \left|(A_{22} - A_{12} A_{11}^{-1} A_{21})_{ij}\right| \leq (k+1) \|A - A_k\|_2 .$$

Because of this relation between row/column-subset selection and the related ideas about picking submatrices of maximum volume, Çivril and Magdon-Ismail [3, 4] looked at the problem of picking a k-subset S of rows of a given matrix $A \in \mathbb{R}^{m \times n}$ such that $\det\left(A_S A_S^T\right)$ is maximized. They show that this problem is NP-hard [3] and moreover, it is NP-hard to even approximate it within a factor of 2^{ck}, for some constant $c > 0$ [4]. This is interesting in the light of our results because we show that even though finding the row-submatrix of maximum volume is NP-hard, we can still sample them with probabilities proportional to their volumes in polynomial time.

2.1 Our results

Our main result is a polynomial time algorithm for exact volume sampling. In Section 4, we give an outline of our Algorithm 1, followed by two possible subroutines given by Algorithms 2 and 3 that could be plugged into it.

Theorem 7 (polynomial-time volume sampling)**.** *The randomized algorithm given by the combination of the algorithm outlined in Algorithm 1 with Algorithm 2 as its subroutine, when given a matrix $A \in \mathbb{R}^{m \times n}$ and an integer $1 \leq k \leq \operatorname{rank}(A)$, outputs a random k-subset of the rows of A according to volume sampling, using $O(kmn^\omega \log n)$ arithmetic operations.*

The basic idea of the algorithm is as follows: instead of picking a k-subset, pick an ordered k-tuple of rows according to volume sampling (i.e., volume sampling suitably extended to all k-tuples such that for any fixed k-subset, all its k! permutations are all equally likely). We observe that the marginal distribution of the first coordinate of such a random tuple can be expressed in terms of coefficients of the characteristic polynomials of AA^T and $B_i B_i^T$, where $B_i \in \mathbb{R}^{m \times n}$ is the matrix obtained by projecting each row of A orthogonal to the i-th row a_i. Using this interpretation, it is easy to sample the first index of the k-tuple with the right marginal probability. Now we project the rows of A orthogonal to the chosen row and repeat to pick the next row, until we have picked k of them.

The algorithm just described informally, if implemented as stated, would have a polynomial dependence in m, n and k, for some low-degree polynomial. We can do better and get a linear dependence in m by working with $A^T A$ in place of AA^T and computing the projected matrices using rank-1 updates (Theorem 7), while still having a polynomial time guarantee and sampling exactly. It would be even faster to perform rank-1 updates to the characteristic polynomial itself, but that requires the computation of the inverse of a polynomial matrix (Proposition 18), and it is not clear to us at this time that there is a fast enough exact algorithm that works for arbitrary matrices. Jeannerod and Villard [14] give an algorithm to invert a *generic* n-by-n matrix with entries of degree d, with n a power of two, in time $O(n^3 d)$. This would lead to the computation of all marginal probabilities for one row in time $O(n^3 + mn^2)$ (a variation of Algorithm 3 and its analysis).

Instead, if we are willing to be more practical, while sacrificing our guarantees, then we can perform rank-1 updates to the characteristic polynomial by using the singular value decomposition (SVD). In [9], an algorithm with cost $O(\min\{mn^2, m^2 n\})$ arithmetic operations is given for the singular value decomposition but the SVD cannot be computed *exactly* and we do not know how its error propagates in our algorithm which uses many such computations. If the SVD of an m-by-n matrix can be computed in time $O(T_{svd})$, this leads to a nearly-exact algorithm for volume sampling in time $O(kT_{svd} + kmn^2)$. See Proposition 18 for details.

Volume sampling was originally defined in [5] to prove Theorem 2, in particular, to show that any matrix A contains k rows in whose span lie the rows of a rank-k approximation to A that is no worse than the best in the Frobenius norm. Efficient volume sampling leads to an efficient selection of k rows that satisfy this guarantee, *in expectation*. In Section 5, we use the method of conditional expectations to derandomize this selection. This gives an efficient deterministic algorithm (Algorithm 4) for row-subset selection with the following guarantee in the Frobenius norm. This guarantee immediately implies a guarantee in the spectral norm, as follows:

Theorem 8 (deterministic row subset selection)**.** *Deterministic Algorithm 4, when given a matrix $A \in \mathbb{R}^{m\times n}$ and an integer $1 \leq k \leq \operatorname{rank}(A)$, outputs a k-subset S of the rows of A, using $O(kmn^\omega \log n)$ arithmetic operations, such that*

$$\begin{aligned}\|A - \pi_S(A)\|_F &\leq \sqrt{k+1}\, \|A - A_k\|_F \\ \|A - \pi_S(A)\|_2 &\leq \sqrt{(k+1)(n-k)}\, \|A - A_k\|_2 \,.\end{aligned}$$

This improves the $O(k\sqrt{\log k})$-approximation of Boutsidis, Drineas and Mahoney [1] for the Frobenius norm case and matches the lower bound shown in Theorem 3 due to [5].

The superlinear dependence on n might be too slow for some applications, while it might be acceptable to perform volume sampling or row/column-subset selection approximately. Our volume sampling algorithm (Algorithm 1) can be made faster, while losing on the exactness, by using the idea of random projection that preserves volumes of subsets. Magen and Zouzias [17] have the following generalization of the Johnson-Lindenstrauss lemma: for m points in $\mathbb{R}^n$ there exists a random projection of them into $\mathbb{R}^d$, where $d = O\left(k^2 \log m/\epsilon^2\right)$, that preserves the volumes of simplices formed by subsets of k or fewer points within $1 \pm \epsilon$. Therefore, we get a $(1 \pm \epsilon)$-approximate volume sampling algorithm that requires $O(mnd)$-time to do the random projection (by matrix multiplication) and then $O(mnd^\omega \log d)$ time for volume sampling on the new m-by-d matrix (according to Theorem 7).

Theorem 9 (fast volume sampling)**.** *Using random projection for dimensionality reduction, the polynomial time algorithm for volume sampling mentioned in Theorem 7 (i.e., Algorithm 1 with Algorithm 2 as its subroutine), gives $(1+\epsilon)$-approximate volume sampling, using*

$$O\left(mn \log m \cdot \frac{k^2}{\epsilon^2} + m \log^\omega m \cdot \frac{k^{2\omega+1}}{\epsilon^{2\omega}} \log(k\epsilon^{-1} \log m) \right).$$

arithmetic operations.

Finally, we show a lower bound for row/column-subset selection in the spectral norm that almost matches our upper bound in terms of the dependence on n.

Theorem 10 (lower bound)**.** *There exists a matrix $A \in \mathbb{R}^{n \times (n+1)}$ such that*

$$\left\| A - \pi_{\{i\}}(A) \right\|_2 = \Omega(\sqrt{n}) \left\| A - A_1 \right\|_2, \quad \text{for all } 1 \leq i \leq n,$$

where $\pi_{\{i\}}(A) \in \mathbb{R}^{n \times (n+1)}$ is the matrix obtained by projecting each row of A onto the span of its i-th row a_i.

3 Preliminaries and notation

For $m \in \mathbb{N}$, let $[m]$ denote the set $\{1, \ldots, m\}$. For any matrix $A \in \mathbb{R}^{m \times n}$, we denote its rows by $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$. For $S \subseteq [m]$, let A_S be the row-submatrix of A given by the rows with indices in S. By $\text{span}\,(S)$ we denote the linear span of $\{a_i \,:\, i \in S\}$ and let $\pi_S(A) \in \mathbb{R}^{m \times n}$ be the matrix obtained by projecting each row of A onto $\text{span}\,(S)$. Hence, $A - \pi_S(A) \in \mathbb{R}^{m \times n}$ is the matrix obtained by projecting each row of A orthogonal to $\text{span}\,(S)$.

Throughout the paper we assume $m \geq n$. This assumption is not needed most of the time, but justifies sometimes working with $A^T A$ instead of AA^T and, more generally, some choices in the design of our algorithms. It is also partially justified by our use of a random projection as a preprocessing step that makes n small.

The singular values of $A \in \mathbb{R}^{m \times n}$ are defined as the positive square-roots of the eigenvalues of $AA^T \in \mathbb{R}^{m \times m}$ (or $A^T A \in \mathbb{R}^{n \times n}$, up to some extra singular values equal to zero), and we denote them by $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m \geq 0$. Well-known identities of the singular values like

$$\text{trace}\left(AA^T\right) = \sum_{i=1}^{m} \sigma_i^2 \quad \text{and} \quad \det\left(AA^T\right) = \prod_{i=1}^{m} \sigma_i^2$$

can be generalized into the following lemma.

Lemma 11. *(Proposition 3.2 in [5]) For any $A \in \mathbb{R}^{m \times n}$,*

$$\sum_{S \subseteq [m] \,:\, |S| = k} \det\left(A_S A_S^T\right) = \sum_{i_1 < \cdots < i_k} \sigma_{i_1}^2 \cdots \sigma_{i_k}^2 = \left| c_{m-k}(AA^T) \right|,$$

where $\sigma_1, \ldots, \sigma_m$ are the singular values of A, i.e., eigenvalues of AA^T, and

$$\det\left(xI - AA^T\right) = x^m + c_{m-1}(AA^T) x^{m-1} + \ldots + c_0(AA^T) = \prod_{i=1}^{m} (x - \sigma_i^2),$$

is the characteristic polynomial of AA^T. *Using* $\det\left(xI - AA^T\right) = x^{m-n}\det\left(xI - A^TA\right)$, *we can alternatively use* $c_{m-k}(AA^T) = c_{n-k}(A^TA)$ *in the above formula, for* $k \leq n$.

Let ω be the exponent of the arithmetic complexity of matrix multiplication. We use that there is an algorithm for computing the characteristic polynomial of an n-by-n matrix using $O(n^\omega \log n)$ arithmetic operations [2, Section 16.6].

Here is another lemma that we will need about dividing determinants into products of two determinants.

Lemma 12. *Let* $A \in \mathbb{R}^{m\times n}$, $S, T \subseteq [m]$, $S \cap T = \emptyset$ *and* $B = A - \pi_S(A)$. *Then*

$$\det\left(A_{S\cup T}A_{S\cup T}^T\right) = \det\left(A_SA_S^T\right)\det\left(B_TB_T^T\right).$$

Proof. Without loss of generality, we can reduce ourselves to the case where $S \cup T$ is all rows of the given matrix: Let $C = A_{S\cup T} \in \mathbb{R}^{|S\cup T|\times n}$, $D = C - \pi_S(C)$. We have $D = B_{S\cup T}$. Then what we want to prove can be rewritten as:

$$\det(CC^T) = \det(C_SC_S^T)\det(D_TD_T^T).$$

To show this, we consider two cases. If $C_SC_S^T$ is singular, then both sides of the equality are zero. If $C_SC_S^T$ is invertible, then we can perform block Gaussian elimination and write

$$\begin{pmatrix} E & F \\ G & H \end{pmatrix}\begin{pmatrix} I & -E^{-1}F \\ 0 & I \end{pmatrix} = \begin{pmatrix} E & 0 \\ G & D - GE^{-1}F \end{pmatrix},$$

applied to $\begin{pmatrix} E & F \\ G & H \end{pmatrix} = C$. Writing the determinants of the block-triangular matrices gives

$$\det(CC^T) = \det(C_SC_S^T)\det(C_TC_T^T - C_TC_S^T(C_SC_S^T)^{-1}C_SC_T^T).$$

Now, the projection of the rows of a matrix K onto the row-space of a matrix L can be written as

$$\pi_L(K) = KL^T(LL^T)^{-1}L,$$

so that $D_T = C_T - C_TC_S^T(C_SC_S^T)^{-1}C_S$, and

$$D_TD_T^T = C_TC_T^T - C_TC_S^T(C_SC_S^T)^{-1}C_SC_T^T.$$

This completes the proof. □

Finally, a well-known lemma about how the determinant of a matrix changes under a rank-1 update.

Lemma 13 (matrix determinant lemma)**.** *For any invertible* $M \in \mathbb{R}^{m\times m}$ *and* $u, v \in \mathbb{R}^m$,

$$\det\left(M + uv^T\right) = (1 + v^TM^{-1}u)\det(M).$$

4 Efficient volume sampling algorithms

We first outline our volume sampling algorithm to convince the reader that volume sampling can be done in polynomial time. In the subsequent subsections, we give improved subroutines to get faster implementations of the same idea.

The main idea behind our algorithm is based on Lemma 14 about the marginal probabilities encountered in volume sampling. To explain this, it is more convenient to look at volume sampling defined as a distribution on k-tuples $(X_1, X_2, \ldots, X_k)$ instead of k-subsets, where each of the $k!$ permutations of a k-subset is equally likely, i.e., for any $(i_1, i_2, \ldots, i_k) \in [m]^k$,

$$\mathsf{Pr}\left(X_1 = i_1, \ldots, X_k = i_k\right) = \begin{cases} \dfrac{\det\left(A_{\{i_1,\ldots,i_k\}} A^T_{\{i_1,\ldots,i_k\}}\right)}{k! \sum_{S \subseteq [m]\,:\,|S|=k} \det\left(A_S A_S^T\right)} & \text{if } i_1, \ldots, i_k \text{ are distinct} \\ \\ 0 & \text{otherwise} \end{cases}$$

Then the marginal probabilities $\mathsf{Pr}\left(X_t = i \mid X_1 = i_1, \ldots, X_{t-1} = i_{t-1}\right)$ have the following interpretation in terms of the coefficients of certain characteristic polynomials.

Lemma 14. *Let $(i_1, \ldots, i_{t-1}) \in [m]^{t-1}$ such that $\mathsf{Pr}\left(X_1 = i_1, \ldots, X_{t-1} = i_{t-1}\right) > 0$, for a random k-tuple $(X_1, X_2, \ldots, X_k)$ from the extended volume sampling over k-tuples. Let $S = \{i_1, \ldots, i_{t-1}\}$, $B = A - \pi_S(A)$ and $C_i = B - \pi_{\{i\}}(B) = A - \pi_{S\cup\{i\}}(A)$. Then,*

$$\mathsf{Pr}\left(X_t = i \mid X_1 = i_1, \ldots, X_{t-1} = i_{t-1}\right) = \frac{\left\|b_i\right\|^2 \left|c_{m-k+t}(C_i C_i^T)\right|}{(k-t+1)\left|c_{m-k+t-1}(BB^T)\right|}.$$

Proof.

$$\begin{aligned}
&\mathsf{Pr}\left(X_t = i \mid X_1 = i_1, \ldots, X_{t-1} = i_{t-1}\right) \\
&= \frac{\sum_{(i_{t+1},\ldots,i_k)\in[m]^{k-t}} \mathsf{Pr}\left(X_1 = i_1, \ldots, X_{t-1} = i_{t-1}, X_t = i, X_{t+1} = i_{t+1}, \ldots, X_k = i_k\right)}{\sum_{(i_t,\ldots,i_k)\in[m]^{k-t+1}} \mathsf{Pr}\left(X_1 = i_1, \ldots, X_{t-1} = i_{t-1}, X_t = i_t, X_{t+1} = i_{t+1}, \ldots, X_k = i_k\right)} \\
&= \frac{(k-t)! \sum_{T\subseteq[m]\,:\,|S\cup\{i\}\cup T|=k, |T|=k-t} \det\left(A_{S\cup\{i\}\cup T} A^T_{S\cup\{i\}\cup T}\right)}{(k-t+1)! \sum_{T\subseteq[m]\,:\,|S\cup T|=k, |T|=k-t+1} \det\left(A_{S\cup T} A^T_{S\cup T}\right)} \\
&= \frac{\sum_{T\subseteq[m]\,:\,|S\cup\{i\}\cup T|=k, |T|=k-t} \det\left(A_S A_S^T\right) \det\left(B_{\{i\}\cup T} B^T_{\{i\}\cup T}\right)}{(k-t+1) \sum_{T\subseteq[m]\,:\,|S\cup T|=k, |T|=k-t+1} \det\left(A_S A_S^T\right) \det\left(B_T B_T^T\right)} \quad \text{by Lemma 12} \\
&= \frac{\sum_{T\subseteq[m]\,:\,|S\cup\{i\}\cup T|=k, |T|=k-t} \left\|b_i\right\|^2 \det\left(C_T C_T^T\right)}{(k-t+1) \sum_{T\subseteq[m]\,:\,|S\cup T|=k, |T|=k-t+1} \det\left(B_T B_T^T\right)} \quad \text{by Lemma 12 applied to } B \\
&= \frac{\left\|b_i\right\|^2 \sum_{T\subseteq[m]\,:\,|T|=k-t} \det\left(C_T C_T^T\right)}{(k-t+1) \sum_{T\subseteq[m]\,:\,|T|=k-t+1} \det\left(B_T B_T^T\right)} \quad \text{since the extra terms in the sum are all zero} \\
&= \frac{\left\|b_i\right\|^2 \left|c_{m-k+t}(C_i C_i^T)\right|}{(k-t+1)\left|c_{m-k+t-1}(BB^T)\right|} \quad \text{by Lemma 11.}
\end{aligned}$$

□

With this lemma in hand, let us consider the following outline of our algorithm. We will later give two more efficient implementations of this outline, depending on how the p_i's are computed.

Algorithm 1. Outline of our volume sampling algorithm

Input: a matrix $A \in \mathbb{R}^{m\times n}$ and $1 \leq k \leq \mathrm{rank}(A)$.
Output: a subset S of k rows of A picked with probability proportional to $\det\left(A_S A_S^T\right)$.

```
1. Initialize $S \leftarrow \emptyset$ and $B \leftarrow A$. For $t = 1$ to $k$ do:
    (a) For $i = 1$ to $m$ compute:
            $p_i = \|b_i\|^2 \cdot \left|c_{m-k+t}(C_i C_i^T)\right|,$
        where $C_i = B - \pi_{\{i\}}(B)$ is a matrix obtained by projecting each row of $B$ orthogonal to $b_i$.
    (b) Pick $i$ with probability proportional to $p_i$. Let $S \leftarrow S \cup \{i\}$ and $B \leftarrow C_i$.
2. Output $S$.
```

Now we show the correctness of the algorithm:

Proposition 15. *The probability that our volume sampling algorithm outlined above picks a k-subset S is proportional to* $\det\left(A_S A_S^T\right)$. *This algorithm can be implemented with a cost of $O(km^3n + km^{\omega+1}\log m)$ arithmetic operations.*

Proof. By Lemma 14, for any $i_1, i_2, \ldots, i_k$ such that $\mathsf{Pr}\left(X_1 = i_1, \ldots, X_k = i_k\right)$, the probability that our algorithm picks a sequence of rows indexed $i_1, i_2, \ldots, i_k$ in that order is equal to

$$\prod_{t=1}^{k} \mathsf{Pr}\left(X_t = i_t \mid X_1 = i_1, \ldots, X_{t-1} = i_{t-1}\right) = \mathsf{Pr}\left(X_1 = i_1, \ldots, X_k = i_k\right) = \frac{\det\left(A_{\{i_1,\ldots,i_k\}} A_{\{i_1,\ldots,i_k\}}^T\right)}{k! \sum_{S\subseteq[m]\,:\,|S|=k} \det\left(A_S A_S^T\right)}.$$

Otherwise, the probability is zero because in the execution of the algorithm, $\|b_i\| = 0$ for some step t. This proves the correctness of our algorithm.

Given that one can compute the characteristic polynomial of an m-by-m matrix in $O(m^\omega \log m)$ (see Section 3), our outline can be implemented with the following count of arithmetic operations: for every t and i, $O(m^2 n)$ to compute $C_i C_i^T$, $O(m^2 n + m^\omega \log m)$ in total for p_i. Thus, volume sampling in $O(km^3 n + km^{\omega+1} \log m)$. □

4.1 Efficient volume sampling without SVD

Here we present the first (faster) subroutine for computing the marginal probabilities p_i's within the volume sampling algorithm outlined in Section 4. The two main ideas behind this subroutine are: (1) We can work with $B^T B, C_i^T C_i \in \mathbb{R}^{n\times n}$ instead of $BB^T, C_i C_i^T \in \mathbb{R}^{m\times m}$. Assuming $m \geq n$, this saves on running time. (2) Each C_i is a rank-1 update of B and therefore, once we have $B^T B$, it can be used to compute all $C_i^T C_i$ efficiently.

Algorithm 2. First subroutine for marginal probabilities

Input: $B \in \mathbb{R}^{m \times n}$.
Output: $p_1, p_2, \ldots, p_m$.

For $i = 1$ to m do:

1. Compute the matrix $C_i^T C_i \in \mathbb{R}^{n \times n}$ by the following formula

$$C_i^T C_i = B^T B - \frac{B^T B b_i b_i^T}{\|b_i\|^2} - \frac{b_i b_i^T B^T B}{\|b_i\|^2} + \frac{b_i b_i^T B^T B b_i b_i^T}{\|b_i\|^4}.$$

2. Compute the characteristic polynomial of $C_i^T C_i$ and output

$$p_i = \|b_i\|^2 \cdot \left| c_{n-k+t}(C_i^T C_i) \right|.$$

Proposition 16. *For any given $B \in \mathbb{R}^{m \times n}$, the Algorithm 2 above computes $p_1, \ldots, p_m$ in $O(mn^\omega \log n)$ arithmetic operations.*

Proof. $B^T B$ can be computed in time $O(mn^2)$. Observe that since C_i is obtained by projecting each row of B orthogonal to b_i,

$$C_i = B - \frac{1}{\|b_i\|^2} B b_i b_i^T,$$

and therefore,

$$C_i^T C_i = B^T B - \frac{B^T B b_i b_i^T}{\|b_i\|^2} - \frac{b_i b_i^T B^T B}{\|b_i\|^2} + \frac{b_i b_i^T B^T B b_i b_i^T}{\|b_i\|^4}.$$

So once we have $B^T B$, for each i, $C_i^T C_i$ can be computed in time $O(n^2)$ and the characteristic polynomial of $C_i^T C_i$ can be computed in time $O(n^\omega \log n)$ [2, Section 16.6]. By Lemma 11, $c_{m-k+t}(C_i C_i^T) = c_{n-k+t}(C_i^T C_i)$ and hence, the above subroutine results into an $O(kmn^\omega \log n)$ time algorithm for volume sampling. □

Theorem 17 (same as Theorem 7)**.** *The randomized algorithm given by the combination of the algorithm outlined in Algorithm 1 with Algorithm 2 as its subroutine, when given a matrix $A \in \mathbb{R}^{m \times n}$ and an integer $1 \leq k \leq \operatorname{rank}(A)$, outputs a random k-subset of the rows of A according to volume sampling, using $O(kmn^\omega \log n)$ arithmetic operations.*

Proof. The proof follows by combining Proposition 15 and Proposition 16, and since we compute all the p_i's simultaneously in each round in $O(mn^\omega \log n)$ arithmetic operations, the total number of arithmetic operations is $O(kmn^\omega \log n)$. □

4.2 Efficient volume sampling using SVD

Taking further the idea that each C_i is a rank-1 update of B, we can give a faster algorithm based on the singular value decomposition of B. Given the singular value decomposition of a matrix and using the matrix determinant lemma (Lemma 13), one can give a precise formula for how the characteristic polynomial changes under a rank-1 update. Using this subroutine in the volume sampling algorithm outlined in Section 4 we get an algorithm for nearly-exact volume sampling (depending on the precision of the computed SVD) in time $O(kT_{svd} + kmn^2)$, where T_{svd} is the running time of SVD on an m-by-n matrix.

Algorithm 3. Second subroutine for marginal probabilities.

Input: $B \in \mathbb{R}^{m \times n}$.
Output: $p_1, p_2, \ldots, p_m$.

1. Compute the (thin) singular value decomposition $B = U\Sigma V^T$, say $U \in \mathbb{R}^{m \times n}$ and $\Sigma, V \in \mathbb{R}^{n \times m}$, and keep the singular values $\sigma_1, \sigma_2, \ldots, \sigma_n$ and define $\sigma_{n+1} = \ldots = \sigma_m = 0$. Also keep the columns of U, i.e., the left singular vectors $u_1, u_2, \ldots, u_n \in \mathbb{R}^m$.

2. Compute the polynomial products

$$\begin{aligned} f(x) &= \prod_{l=1}^{m} (x - \sigma_l^2), \quad \text{and} \\ g_j(x) &= \prod_{l \neq j} (x - \sigma_l^2), \quad \text{for all } 1 \leq j \leq m. \end{aligned}$$

3. For $i = 1$ to m output:

$$p_i = \|b_i\|^2 \cdot \left| \text{coefficient of } x^{m-k+t} \text{ in } f(x) + \frac{1}{\|b_i\|^2} \sum_{j=1}^{n} \sigma_j^2 (u_j)_i^2 g_j(x) \right|.$$

Proposition 18. *In the real arithmetic model and given exact U and Σ, using the Algorithm 3 as a subroutine inside Algorithm 1 outlined for volume sampling, we get an algorithm for volume sampling. If T_{svd} is the running time for computing the singular value decomposition of m-by-n matrices, the algorithm runs in time $O(kT_{svd} + kmn^2)$.*

Proof. Using the matrix determinant lemma (Lemma 13), the characteristic polynomial of $C_i C_i^T$

can be written as

$$
\begin{aligned}
\det\left(xI - C_iC_i^T\right) &= \det\left(xI - BB^T + \frac{1}{\|b_i\|^2}(Bb_i)(Bb_i)^T\right) \\
&= \left(1 + \frac{1}{\|b_i\|^2} b_i^T B^T (xI - BB^T)^{-1} Bb_i\right) \det\left(xI - BB^T\right) \\
&= \left(1 + \frac{1}{\|b_i\|^2} b_i^T B^T (xI - \tilde{U}\tilde{\Sigma}^2\tilde{U}^T)^{-1} Bb_i\right) \det\left(xI - BB^T\right) \\
&\quad \text{by extending } U, \Sigma, V \text{ to get } B = \tilde{U}\tilde{\Sigma}\tilde{V}^T \text{ with } \tilde{U}, \tilde{\Sigma} \in \mathbb{R}^{m \times m} \text{ and } \tilde{V} \in \mathbb{R}^{m \times n} \\
&= \left(1 + \frac{1}{\|b_i\|^2} b_i^T B^T \tilde{U} (xI - \tilde{\Sigma}^2)^{-1} \tilde{U}^T Bb_i\right) \det\left(xI - BB^T\right) \\
&= \left(1 + \frac{1}{\|b_i\|^2} b_i^T \tilde{V}\tilde{\Sigma}^T (xI - \tilde{\Sigma}^2)^{-1} \tilde{\Sigma}\tilde{V}^T b_i\right) \det\left(xI - BB^T\right) \\
&= \left(1 + \frac{1}{\|b_i\|^2} \sum_{j=1}^{m} \frac{\sigma_j^2 (\tilde{u}_j)_i^2}{x - \sigma_j^2}\right) \prod_{l=1}^{m} (x - \sigma_l^2) \\
&= \left(1 + \frac{1}{\|b_i\|^2} \sum_{j=1}^{n} \frac{\sigma_j^2 (u_j)_i^2}{x - \sigma_j^2}\right) \prod_{l=1}^{m} (x - \sigma_l^2) \\
&= \prod_{l=1}^{m} (x - \sigma_l^2) + \frac{1}{\|b_i\|^2} \sum_{j=1}^{n} \sigma_j^2 (u_j)_i^2 \prod_{l \neq j} (x - \sigma_l^2) \\
&= f(x) + \frac{1}{\|b_i\|^2} \sum_{j=1}^{n} \sigma_j^2 (u_j)_i^2 g_j(x).
\end{aligned}
$$

Thus,

$$
c_{m-k+t}(C_iC_i^T) = \text{coefficient of } x^{m-k+t} \text{ in } f(x) + \frac{1}{\|b_i\|^2} \sum_{j=1}^{n} \sigma_j^2 (u_j)_i^2 g_j(x).
$$

Once we have the singular value decomposition of B, $f(x)$ and $g_j(x)$ can all be computed in time $O(n^2)$ using polynomial products. This is because there are at most n non-zero σ_i's. Thus, $f(x)$ and all the $g_j(x)$ for $1 \leq j \leq m$ can be computed in time $O(mn^2)$ and then using the above formula we get $c_{m-k+t}(C_iC_i^T)$. □

4.3 Approximate volume sampling in nearly linear time

Magen and Zouzias [17] showed that the random projection lemma of Johnson and Lindenstrauss can be generalized to preserve volumes of subsets after embedding. Here is a restatement of Theorem 1 of [17] using $O(\epsilon/k)$ instead of ϵ in their original statement.

Theorem 19. *[17] For any $A \in \mathbb{R}^{m \times n}$, $1 \leq k \leq n$ and $0 < \epsilon \leq 1/2$, there is*

$$
d = O\left(\frac{k^2 \log m}{\epsilon^2}\right),
$$

and there is a mapping $f : \mathbb{R}^n \to \mathbb{R}^d$ such that

$$\det\left(A_S A_S^T\right) \leq \det\left(\tilde{A}_S \tilde{A}_S^T\right) \leq (1+\epsilon)\det\left(A_S A_S^T\right),$$

for all $S \subseteq [m]$ such that $|S| \leq k$, where $\tilde{A} \in \mathbb{R}^{m \times d}$ has its i-th row as $f(a_i)$. Moreover, f is a linear mapping given by multiplication with a random n by d matrix with i.i.d. Gaussian entries, so computing $\tilde{A}$ takes time $O(mnd)$.

Theorem 20 (same as Theorem 9)**.** *Using random projection for dimensionality reduction, the polynomial time algorithm for volume sampling mentioned in Theorem 7 (i.e., Algorithm 1 with Algorithm 2 as its subroutine), gives $(1+\epsilon)$-approximate volume sampling, using*

$$O\left(mn\log m \cdot \frac{k^2}{\epsilon^2} + m\log^\omega m \cdot \frac{k^{2\omega+1}}{\epsilon^{2\omega}} \log(k\epsilon^{-1}\log m)\right).$$

arithmetic operations.

Proof. Using Theorem 19 and doing volume sampling of k-subsets of rows from $\tilde{A}$ gives $(1+\epsilon)$-approximation to the volume sampling of k-subsets of rows from A. This can be done in two steps: first, we compute $\tilde{A}$ using matrix multiplication in time $O(mnd)$ and second, we do volume sampling on $\tilde{A}$ using the algorithm from Subsection 4.1. Overall, it takes time $O(mnd + kmd^\omega \log d)$, which is equal to

$$O\left(mn\log m \cdot \frac{k^2}{\epsilon^2} + m\log^\omega m \cdot \frac{k^{2\omega+1}}{\epsilon^{2\omega}}\right).$$

Moreover, this can be implemented using only one pass over the matrix A with extra space $m\log m \cdot k^2/\epsilon^2$. □

5 Derandomized row/column-subset selection

Our derandomized row-subset selection algorithm is based on a derandomization of the volume sampling algorithm in Section 4, using the method of conditional expectations. Again, it may be easier to consider volume sampling extended to random k-tuples $(X_1, \ldots, X_k)$ where

$$\mathsf{Pr}\left(X_1 = i_1, \ldots, X_k = i_k\right) = \begin{cases} \dfrac{\det\left(A_{\{i_1,\ldots,i_k\}} A_{\{i_1,\ldots,i_k\}}^T\right)}{k! \sum_{S\subseteq[m]\,:\,|S|=k} \det\left(A_S A_S^T\right)} & \text{if } i_1,\ldots,i_k \text{ are distinct} \\ \\ 0 & \text{otherwise} \end{cases}$$

From Theorem 2 we know that

$$\mathsf{E}\left[\left\|A - \pi_{\{X_1,\ldots,X_k\}}(A)\right\|_F^2\right] \leq (k+1)\left\|A - A_k\right\|_F^2,$$

where the expectation is over $(X_1, \ldots, X_k)$.

Let us consider $i_1, \ldots, i_{t-1}$ for which $\mathsf{Pr}\left(X_1 = i_1, \ldots, X_{t-1} = i_{t-1}\right) > 0$. Let $S = \{i_1, \ldots, i_{t-1}\}$ and look at the conditional expectation. The following lemma shows that these conditional expectations have an easy interpretation in terms of the coefficients of certain characteristic polynomials, and hence can be computed efficiently.

Lemma 21. *Let* $(i_1,\ldots,i_{t-1}) \in [m]^{t-1}$ *be such that* $\mathsf{Pr}\left(X_1 = i_1,\ldots,X_{t-1} = i_{t-1}\right) > 0$ *for a random* k*-tuple* $(X_1, X_2, \ldots, X_k)$ *from extended volume sampling. Let* $S = \{i_1,\ldots,i_{t-1}\}$ *and* $B = A - \pi_S(A)$. *Then*

$$\mathsf{E}\left[\left\|A - \pi_{\{X_1,\ldots,X_k\}}(A)\right\|_F^2 \mid X_1 = i_1,\ldots,X_{t-1} = i_{t-1}\right] = \frac{(k-t+2)c_{m-k+t-2}(BB^T)}{c_{m-k+t-1}(BB^T)}.$$

Proof.

$$\begin{aligned}
&\mathsf{E}\left[\left\|A - \pi_{\{X_1,\ldots,X_k\}}(A)\right\|_F^2 \mid X_1 = i_1,\ldots,X_{t-1} = i_{t-1}\right] \\
&= \sum_{(i_t,\ldots,i_k)\in[m]^{k-t+1}} \left\|A - \pi_{\{i_1,\ldots,i_k\}}(A)\right\|_F^2 \mathsf{Pr}\left(X_1 = i_1,\ldots,X_k = i_k \mid X_1 = i_1,\ldots,X_{t-1} = i_{t-1}\right) \\
&= \sum_{(i_t,\ldots,i_k)\in[m]^{k-t+1}} \left\|A - \pi_{\{i_1,\ldots,i_k\}}(A)\right\|_F^2 \frac{\mathsf{Pr}\left(X_1 = i_1,\ldots,X_k = i_k\right)}{\mathsf{Pr}\left(X_1 = i_1,\ldots,X_{t-1} = i_{t-1}\right)} \\
&= \sum_{(i_t,\ldots,i_k)\in[m]^{k-t+1}} \frac{\sum_{l=1}^m \|d_l\|^2 \det\left(A_{\{i_1,\ldots,i_k\}}A_{\{i_1,\ldots,i_k\}}^T\right)}{\sum_{(j_t,\ldots,j_k)\in[m]^{k-t+1}} \det\left(A_{\{i_1,\ldots,i_{t-1},j_t,\ldots,j_k\}}A_{\{i_1,\ldots,i_{t-1},j_t,\ldots,j_k\}}^T\right)} \\
&\quad \text{where } D = A - \pi_{\{i_1,\ldots,i_k\}}(A) \\
&= \frac{\sum_{(i_t,\ldots,i_k)\in[m]^{k-t+1}} \sum_{l\notin\{i_1,\ldots,i_k\}} \det\left(A_{\{l,i_1,\ldots,i_k\}}A_{\{l,i_1,\ldots,i_k\}}^T\right)}{\sum_{(j_t,\ldots,j_k)\in[m]^{k-t+1}} \det\left(A_{\{i_1,\ldots,i_{t-1},j_t,\ldots,j_k\}}A_{\{i_1,\ldots,i_{t-1},j_t,\ldots,j_k\}}^T\right)} \\
&= \frac{(k-t+1)!\sum_{T\subseteq[m]\,:\,|S\cup T|=k,|T|=k-t+1} \sum_{l\notin S\cup T} \det\left(A_{\{l\}\cup S\cup T}A_{\{l\}\cup S\cup T}^T\right)}{(k-t+1)!\sum_{T\subseteq[m]\,:\,|S\cup T|=k,|T|=k-t+1} \det\left(A_{S\cup T}A_{S\cup T}^T\right)} \\
&= \frac{\sum_{T\subseteq[m]\,:\,|S\cup T|=k,|T|=k-t+1} \sum_{l\notin S\cup T} \det\left(A_S A_S^T\right)\det\left(B_{\{l\}\cup T}B_{\{l\}\cup T}^T\right)}{\sum_{T\subseteq[m]\,:\,|S\cup T|=k,|T|=k-t+1} \det\left(A_S A_S^T\right)\det\left(B_T B_T^T\right)} \quad \text{by Lemma 12} \\
&= \frac{\sum_{T\subseteq[m]\,:\,|S\cup T|=k,|T|=k-t+1} \sum_{l\notin S\cup T} \det\left(B_{\{l\}\cup T}B_{\{l\}\cup T}^T\right)}{\sum_{T\subseteq[m]\,:\,|S\cup T|=k,|T|=k-t+1} \det\left(B_T B_T^T\right)} \\
&= \frac{(k-t+2)\sum_{T\subseteq[m]\,:\,|S\cup T|=k+1,|T|=k-t+2} \det\left(B_T B_T^T\right)}{\sum_{T\subseteq[m]\,:\,|S\cup T|=k,|T|=k-t+1} \det\left(B_T B_T^T\right)} \\
&= \frac{(k-t+2)\sum_{T\subseteq[m]\,:\,|T|=k-t+2} \det\left(B_T B_T^T\right)}{\sum_{T\subseteq[m]\,:\,|T|=k-t+1} \det\left(B_T B_T^T\right)} \quad \text{the extra terms in the numerator and the denominator are zero} \\
&= \frac{(k-t+2)\left|c_{m-k+t-2}(BB^T)\right|}{\left|c_{m-k+t-1}(BB^T)\right|} \quad \text{by Lemma 11.}
\end{aligned}$$

□

Knowing the above lemma, it is easy to derandomize our algorithm outlined for volume sampling. In each step, we just compute the new conditional expectations for each additional i, and finally pick the i that minimizes the conditional expectation.

Algorithm 4. Derandomized row/column-subset selection

```
Input: a matrix $A \in \mathbb{R}^{m \times n}$ and $1 \leq k \leq \operatorname{rank}(A)$.
Output: a subset $S$ of $k$ rows of $A$ with the guarantee

    $\|A - \pi_S(A)\|_F^2 \leq (k+1)\|A - A_k\|_F^2 .$

  1. Initialize $S \leftarrow \emptyset$ and $B \leftarrow A$. For $t = 1$ to $k$ do:
     (a) For $i = 1$ to $m$ do: compute $c_{n-k+t-1}(C_i^T C_i)$ and $c_{n-k+t}(C_i^T C_i)$, where $C_i = B - \pi_{\{i\}}(B)$ is the matrix obtained by projecting each row of $B$ orthogonal to $b_i$.
     (b) Pick $i$ that minimizes $\left|c_{n-k+t-1}(C_i^T C_i)\right| / \left|c_{n-k+t}(C_i^T C_i)\right|$. Let $S \leftarrow S \cup \{i\}$ and $B \leftarrow C_i$.
  2. Output $S$.
```

Theorem 22 (same as Theorem 8)**.** *Deterministic Algorithm 4, when given a matrix $A \in \mathbb{R}^{m \times n}$ and an integer $1 \leq k \leq \operatorname{rank}(A)$, outputs a k-subset S of the rows of A, using $O(kmn^\omega \log n)$ arithmetic operations, such that*

$$\|A - \pi_S(A)\|_F \leq \sqrt{k+1}\, \|A - A_k\|_F$$
$$\|A - \pi_S(A)\|_2 \leq \sqrt{(k+1)(n-k)}\, \|A - A_k\|_2 .$$

Proof. By applying Lemma 21 to $S \cup \{i\}$ instead of S, as $C_i = B - \pi_{\{i\}}(B) = A - \pi_{S \cup \{i\}}(A)$, we see that the step t of our algorithm picks i that minimizes

$$\mathsf{E}\left[\left\|A - \pi_{\{X_1,\ldots,X_k\}}(A)\right\|_F^2 \mid X_1 = i_1, \ldots, X_{t-1} = i_{t-1}, X_t = i\right] = \frac{(k-t+1)\left|c_{n-k+t-1}(C_i^T C_i)\right|}{\left|c_{n-k+t}(C_i^T C_i)\right|} = \frac{(k-t+1)\left|c_{m-k+t-1}(C_i C_i^T)\right|}{\left|c_{m-k+t}(C_i C_i^T)\right|}.$$

The correctness of our algorithm follows immediately from observing that in each step t,

$$\mathsf{E}\left[\left\|A - \pi_{\{X_1,\ldots,X_k\}}(A)\right\|_F^2 \mid X_1 = i_1, \ldots, X_{t-1} = i_{t-1}\right]$$
$$= \sum_{i=1}^{m} \mathsf{Pr}\left(X_t = i \mid X_1 = i_1, \ldots, X_{t-1} = i_{t-1}\right) \mathsf{E}\left[\left\|A - \pi_{\{X_1,\ldots,X_k\}}(A)\right\|_F^2 \mid X_1 = i_1, \ldots, X_{t-1} = i_{t-1}, X_t = i\right]$$

and that we started with

$$\mathsf{E}\left[\left\|A - \pi_{\{X_1,\ldots,X_k\}}(A)\right\|_F^2\right] \leq (k+1)\|A - A_k\|_F^2 .$$

The guarantee for spectral norm follows immediately from our guarantee in the Frobenius norm, just using properties of norms and the fact that $\operatorname{rank}(A - A_k) \leq n - k$:

$$\|A - \pi_S(A)\|_2^2 \leq \|A - \pi_S(A)\|_F^2 \leq (k+1)\|A - A_k\|_F^2 \leq (k+1)(n-k)\|A - A_k\|_2^2.$$

Moreover, this algorithm runs in time $O(kmn^\omega \log n)$ if we use the subroutine in Subsection 4.1 to compute the characteristic polynomial of $C_i C_i^T$ using that of $C_i^T C_i$. □

6 Lower bound for rank-1 spectral approximation using one row

Here we show a lower bound for row/column-subset selection. We prove that there is a matrix $A \in \mathbb{R}^{n\times(n+1)}$ such that using the span of any single row of it, we can get only $\Omega(\sqrt{n})$-approximation in the spectral norm for the nearest rank-1 matrix to A. This can be generalized to a similar $\Omega(\sqrt{n})$ lower bound for general k by using a matrix with k block-diagonal copies of A.

Theorem 23 (same as Theorem 10)**.** *There exists a matrix $A \in \mathbb{R}^{n\times(n+1)}$ such that*

$$\left\|A - \pi_{\{i\}}(A)\right\|_2 = \Omega(\sqrt{n}) \left\|A - A_1\right\|_2, \quad \textit{for all } 1 \le i \le n,$$

where $\pi_{\{i\}}(A) \in \mathbb{R}^{n\times(n+1)}$ is the matrix obtained by projecting each row of A onto the span of its i-th row a_i.

Proof. Consider $A \in \mathbb{R}^{n\times(n+1)}$ with entries as follows:

$$\begin{pmatrix} 1 & \epsilon & 0 & \dots & 0 \\ 1 & 0 & \epsilon & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 & \epsilon \end{pmatrix}, \qquad 0 < \epsilon < 1.$$

Let B be the best rank-1 approximation to A whose rows lie in the span of $(1, \epsilon, 0, \dots, 0)$ (or for that matter, any fixed row of A). Then, we want to show that

$$\|A - B\|_2 \ge \frac{\sqrt{n}}{2} \|A - A_1\|_2 = \frac{\sqrt{n}}{2} \sigma_2(A).$$

We first compute the singular values of A, i.e., the positive square roots of the eigenvalue of $AA^T \in \mathbb{R}^{n\times n}$.

$$AA^T = \begin{pmatrix} 1+\epsilon^2 & 1 & 1 & \dots & 1 \\ 1 & 1+\epsilon^2 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 & 1 \\ 1 & \dots & 1 & 1+\epsilon^2 & 1 \\ 1 & \dots & 1 & 1 & 1+\epsilon^2 \end{pmatrix}.$$

$(1, 1, \dots, 1)$ is an eigenvector of AA^T with eigenvalue $n + \epsilon^2$. Thus, $\sigma_1(A) = \sqrt{n + \epsilon^2}$. Observe that, by symmetry, all other singular values of A must be equal, i.e., $\sigma_2(A) = \sigma_3(A) = \ldots = \sigma_n(A)$. However,

$$\|A\|_F^2 = \sum_{ij} A_{ij}^2 = n + n\epsilon^2 = \sum_{i=1}^{n} \sigma_i(A)^2 = \sigma_1(A)^2 + (n-1)\sigma_2(A)^2 = n + \epsilon^2 + (n-1)\sigma_2(A)^2.$$

Therefore, $\|A - A_1\|_2 = \sigma_2(A) = \epsilon$.

Now denote the i-th row of A by a_i. By definition, the i-th row of B is the projection of a_i onto span (a_1). We are interested in the singular values of $A - B$. For $i \ge 2$:

$$\begin{aligned} a_i - b_i &= a_i - \frac{\langle a_i, a_1\rangle}{\|a_1\|^2} a_1 \\ &= \left(\frac{\epsilon^2}{1+\epsilon^2}, \frac{-\epsilon}{1+\epsilon^2}, 0, \underbrace{\qquad\epsilon\qquad}_{(i+1)\text{-th coord.}}, 0, \dots, 0 \right). \end{aligned}$$

Thus, $(A-B)(A-B)^T \in \mathbb{R}^{n\times n}$ can be written as

$$(A-B)(A-B)^T = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & \frac{\epsilon^2(2+\epsilon^2)}{1+\epsilon^2} & \frac{\epsilon^2}{1+\epsilon^2} & \dots & \frac{\epsilon^2}{1+\epsilon^2} \\ \dots & \frac{\epsilon^2}{1+\epsilon^2} & \frac{\epsilon^2(2+\epsilon^2)}{1+\epsilon^2} & \dots & \frac{\epsilon^2}{1+\epsilon^2} \\ 0 & \frac{\epsilon^2}{1+\epsilon^2} & \dots & \frac{\epsilon^2(2+\epsilon^2)}{1+\epsilon^2} & \frac{\epsilon^2}{1+\epsilon^2} \\ 0 & \frac{\epsilon^2}{1+\epsilon^2} & \dots & \frac{\epsilon^2}{1+\epsilon^2} & \frac{\epsilon^2(2+\epsilon^2)}{1+\epsilon^2} \end{pmatrix}.$$

Again, $(0,1,1,\dots,1)$ is the top eigenvector of $(A-B)(A-B)^T$ and using this we get,

$$\|A-B\|_2^2 = \sigma_1(A-B)^2 = \frac{\epsilon^2(2+\epsilon^2)}{1+\epsilon^2} + (n-2)\frac{\epsilon^2}{1+\epsilon^2}.$$

Therefore,

$$\|A-B\|_2 = \frac{\epsilon}{\sqrt{1+\epsilon^2}}\sqrt{n+\epsilon^2} \geq \frac{\sqrt{n}}{2}\|A-A_1\|_2 .$$

$\square$

7 Discussion

We analyzed efficient algorithms for volume sampling that can be used for row/column subset selection. Here are some ideas for future investigation suggested by this work:

- It would be interesting to explore how these algorithmic ideas are related to determinantal sampling [16, 13] and, in particular, the generation of random spanning trees.
- Find practical counterparts of the algorithms discussed here. In particular, we do not analyze the numerical stability of our algorithms.
- Is there an efficient algorithm for volume sampling based on random walks? This question is inspired by MCMC as well as random walk algorithms for the generation of random spanning trees.

References

[1] C. Boutsidis, P. Drineas, and M. Mahoney. An improved approximation algorithm for the column subset selection problem. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2009.

[2] P. Bürgisser, M. Clausen, and M. A. Shokrollahi. *Algebraic complexity theory*, volume 315 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1997. With the collaboration of Thomas Lickteig.

[3] A. Çivril and M. Magdon-Ismail. On selecting the maximum volume submatrix of a matrix and related problems. *Theoretical Computer Science*, 410 (47-49):4801–4811, 2009.

[4] A. Çivril and M. Magdon-Ismail. Exponential inapproximability of selecting a maximum volume submatrix. unpublished manuscript, 2010.

[5] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. *Theory of Computing*, 2(1):225–247, 2006.

[6] A. Deshpande and S. Vempala. Adaptive sampling and fast low-rank matrix approximation. In *International Workshop on Randomization and Computation (RANDOM)*, 2006.

[7] P. Drineas, M. Mahoney, and S. Muthukrishnan. Relative-error cur matrix decompositions. *SIAM Journal of Matrix Analysis and Applications*, 30 (2):844–881, 2008.

[8] A. Frieze, R. Kannan, and S. Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *Journal of the ACM (JACM)*, 51(6):1025–1041, 2004.

[9] G. Golub and C. van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.

[10] S. Goreinov and E. Tyrtyshnikov. The maximum-volume concept in approximation by low-rank matrices. *Contemporary Mathematics*, 280:47–51, 2001.

[11] S. Goreinov, E. Tyrtyshnikov, and N. Zamarashkin. Pseudo-skeleton approximations by matrices of maximal volume. *Mathematicheskie Zametki*, 62:619–623, 1997.

[12] M. Gu and S. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal of Scientific Computing*, 17:848–869, 1996.

[13] J. Hough, M. Krishnapur, Y. Peres, and B. Virág. Determinantal processes and independence. *Probability Surveys*, 3:206–229, 2006.

[14] C. Jeannerod and G. Villard. Essentially optimal computation of the inverses of generic polynomial matrices. *Journal of Complexity*, 21 (1):72–86, 2005.

[15] R. Kannan and S. Vempala. Spectral algorithms. *Foundations and Trends in Theoretical Computer Science*, 4:157–288, 2009.

[16] R. Lyons. Determinantal probability measures. *Publications Mathématiques de l'IHÉS*, 98(1):167–212, 2003.

[17] A. Magen and A. Zouzias. Near optimal dimensionality reductions that preserve volumes. In *International Workshop on Randomization and Computation (RANDOM)*, 2008.

[18] M. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences USA*, 106:697–702, 2009.

[19] C.-T. Pan. On the existence and computation of rank-revealing LU factorizations. *Linear Algebra and its Applications*, 316 (1-3):199–222, 2000.